ENHANCING SPEECH INTELLIGIBILITY USING VARIABLE-RATE TIME-SCALE MODIFICATION

FIELD OF THE INVENTION

[01] The present invention relates to a modification of a speech signal in order to enhance the intelligibility of the associated speech.

BACKGROUND OF THE INVENTION

[02] Reducing the bandwidth associated with a speech signal for coding applications often results in the listener having difficulty in understanding consonant sounds. It is desirable to strengthen the available acoustic cues to make consonant contrasts more distinct, and potentially more robust to subsequent coding degradations. The intelligibility of speech is an important issue in the design of speech coding algorithms. In narrowband speech the distinction between consonants can be poor, even in quiet conditions and prior to signal encoding. This happens most often for those consonants that differ by place of articulation. While reduced intelligibility may be partly attributed to the removal of high frequency information, resulting in a loss of cue redundancy, the problem is often intensified by the weak nature of the acoustic cues available in consonants. It is thus advantageous to strengthen the identifying cues to improve speech perception.

[03] Speakers naturally revise their speech when talking to impaired listeners or in adverse environments. This type of speech, known as clear speech, is typically half the speaking rate of conversational speech. Other differences include longer formant transitions, more salient consonant contrasts (increased consonant–vowel ratio, CVR), and pauses, which are more frequent and longer in duration. Prior art attempts to improve intelligibility involve artificially modifying speech to possess these characteristics. Although increased CVR may lead to improved intelligibility in the presence of noise due to the inherent low energy of consonants, in a noise-free environment, significantly modifying the natural relative CV amplitudes of a phoneme can prove unfavorable by creating the perception of a different phoneme.

[04] Techniques for the selective modification of speech duration to improve or maintain the level of intelligibility have also been proposed. There are two main approaches. The first approach modifies the speech only during steady-state sections by increasing the speaking rate without causing a corresponding decrease in quality or intelligibility. Alternatively, the speech may be modified only during non-steady-state, transient regions. Both approaches result in a change in the signal duration, and both detect and treat transient regions of speech in a different manner from the rest of the signal. For real-time applications, however, the signal duration must remain essentially unchanged.

[05] Thus, there is a need to enhance the intelligibility of narrowband speech without lengthening the overall duration of the signal.

BRIEF SUMMARY OF THE INVENTION

[06] Transmission and processing of a speech signal is often associated with bandwidth reduction, packet loss, and the exacerbation of noise. These degradations can result in a corresponding increase of consonant confusions for speech applications. Strengthening the available acoustic cues to make consonant contrasts more distinct may provide greater robustness to subsequent coding degradations. The present invention provides methods for enhancing speech intelligibility using variable rate time-scale modification of a speech signal. Frequency domain characteristics of an input speech signal are modified to produce an intermediate speech signal, such that acoustic cues of the input speech signal are enhanced. Time domain characteristics of the intermediate speech signal are then modified to produce an output signal, such that steady-state and non-steady-state parts of the intermediate speech signal of the intermediate speech signal are oppositely modified.

[07] An exemplary embodiment is disclosed that enhances the intelligibility of narrowband speech without lengthening the overall duration of the signal. The invention incorporates both spectral enhancements and variable-rate time-scaling procedures to improve the salience of initial consonants, particularly the perceptually important formant transitions. Emphasis is transferred from the dominating vowel to the preceding consonant through adaptation of the phoneme timing structure.

[08] In a second exemplary embodiment of the present invention, the technique is applied as a preprocessor to the Mixed Excitation Linear Prediction (MELP) coder. The technique is thus adapted to produce a signal with qualities favorable for MELP encoding. Variations of the embodiment can be applied to other types of speech coders, including code excited linear prediction (CELP), vector sum excitation (VSELP), waveform interpolation (WI), multiband excitation (MBE), linear prediction coding (LPC), pulse code modulation (PCM), differential pulse code modulation (DPCM), and adaptive differential pulse code modulation (ADPCM).

BRIEF DESCRIPTION OF THE DRAWINGS

[09] Figure 1 is a block diagram of the enhancement algorithm of the present invention;

[10] Figure 2 depicts a time-scale modification syllable (TSMS) of the word "sank";

[11] Figure 3 depicts measures used to locate syllables to time-scale;

[12] Figure 4 depicts locating the time-scale modification syllable for the word "fin" according to a speech waveform;

[13] Figure 5 depicts locating the time-scale modification syllable for the word "fin" according to an energy contour;

[14] Figure 6 depicts locating the time-scale modification syllable for the word "fin" according to a spectral feature transition rate (SFTR);

[15] Figure 7 is a block diagram of the variable-rate time-scale modification procedure;

[16] Figure 8 is a flow diagram corresponding to Figure 7;

[17] Figure 9 depicts an input signal corresponding to the word "fin";

[18] Figure 10 depicts the self- determined scaling factors during the time duration corresponding to FIG. 9;

[19] Figure 11 depicts the total delay (including a 100ms look-ahead delay) during the time duration corresponding to FIG. 9;

[20] Figure 12 depicts the output variable rate time-scale modification of the word "fin";

[21] Figure 13 depicts the effect of WSOLA pitch errors on the MELP coded signal having a time scale modification (TSM) signal with single "best-match"/pitch error;

[22] Figure 14 depicts the effect of WSOLA pitch errors on a MELP coded signal having a the MELP coded enhanced signal;

[23] Figure 15 depicts an intelligibility enhancement pre-processor for a MELPe speech coder; and

[24] Figure 16 is a flow diagram corresponding to Figure 15.

DETAILED DESCRIPTION OF THE INVENTION

[25] The vowel sounds (often referenced as voiced speech) carry the power in speech, but the consonant sounds (often referenced as unvoiced speech) are the most important for understanding. However, consonants, especially those within the same class, are often difficult to differentiate and are more vulnerable to many forms of signal degradation. For example, speech (as conveyed by a signal) may be degraded in a telecommunications network that is characterized by packet loss (for a packetized signal) or by noise. By appropriately processing the speech signal, the processed speech signal may be more immune to subsequent degradations.

[26] Preliminary experiments analyzing the distinction between confusable word pairs show that intelligibility can be improved if the test stimuli were presented twice to the listener, as opposed to only once. It is hypothesized that when the first time the word is heard, the high-intensity, longer duration vowel partially masks the adjacent consonant. When the word is repeated, the vowel is already known and expected, allowing the listener to then focus on identifying the consonant. To eliminate the need for repetition, it is desirable to reduce the vowel emphasis, and increase the salience of the consonant cues to weaken the masking effect.

[27] The most confusable consonant pairs are those that differ by place of articulation, e.g. /p/-/t/, /f/-/th/. These contain their main distinctive feature during their co-articulation

with adjacent phonemes, characterized by the consonant-vowel formant transitions. To emphasize the formant structure, transient regions of speech are slowed down, while the contrasts are increased between spectral peaks and valleys. In addition, the steady state vowel following a syllable-initial consonant is compressed. The compression serves at least three main purposes. First, it accentuates the longer consonant length; second, it preserves the waveform rhythm to maintain naturalness; and third, it results in minimum overall phrase time duration change, which allows the technique of the present invention to be employed in real-time applications.

[28] Common methods used to modify the time duration of speech without altering perceived frequency attributes are overlap-add (OLA) techniques. OLA is a time-domain technique that modifies the time-scale of a signal without altering its perceived frequency attributes. OLA constructs a modified signal that has a short-time Fourier Transform (STFT) maximally close to that of the original signal. These techniques are popular due to their low complexity, allowing for real-time implementation. OLA techniques average overlapping frames of a signal at points of highest correlation to obtain a time-scaled signal, which maintains the local pitch and spectral properties of the original signal. To reduce discontinuities at waveform boundaries and improve synchronization, the waveform similarity overlap-add (WSOLA) technique was developed. WSOLA overcomes distortions of OLA by selecting the segment for overlap-addition, within a given tolerance of the target position, such that the synthesized waveform has maximal similarity to the original signal across segment boundaries. The synthesis equation for WSOLA with regularly spaced synthesis instants kL and a symmetric unity gain window, $v(n)$, is:

$$y(n) = \sum_k v(n - kL).x(n + \tau^{-1}(kL) + \Delta_k - kL) \qquad (1)$$

where $\tau^{-1}(kL)$ represents time instants on the input signal, and $\Delta_k \in [-\Delta_{max} \ldots \Delta_{max}]$ is the tolerance introduced to achieve synchronization.

[29] To find the position of the best-matched segment, the normalized cross-correlation function is maximized as follows:

$$c_n(m,\delta)=\frac{\sum_{n=0}^{N-1}x(n+\tau^{-1}((m-1)L)+\Delta_{m-1}+L).x(\sum_{n=0}^{N-1}x(n+\tau^{-1}(mL)+\delta)}{\sqrt{\sum_{n=0}^{N-1}x^2(n+\tau^{-1}(mL)+\delta)}} \quad (2)$$

where N is the window length.

[30] With the present invention, the intelligibility enhancement algorithm enhances the identifying features of syllable-initial consonants. It focuses mainly on improving the distinctions between initial consonants that differ by place of articulation, i.e. consonants within the same class that are produced at different points of the vocal tract. These are distinguished primarily by the location and transition of the formant frequencies. The method can be viewed as a redistribution of segment durations at a phonetic level, combined with frequency-selective amplification of acoustic cues. This emphasizes the co-articulation between a consonant and its following vowel. In one embodiment the algorithm is used in a preprocessor in real-time speech applications. The enhancement strategy, illustrated in Figure 1, is divided into two main parts: a first portion 101 for modification of frequency domain characteristics, and a second portion 102 for modification of time-domain characteristics.

[31] In the exemplary embodiment of the present invention, modification of the frequency domain characteristics in first portion 101 involves adaptive spectral enhancement (enhancement filter 103) to make the spectral peaks more distinct, and emphasis (tilt compensator 104) of the higher frequencies to reduce the upward spread of masking. This is then followed by the time-domain modification of second portion 102, which automatically identifies the segments to be modified (syllable segmentation 105), determines the appropriate time-scaling factor (scaling factor determination 106) for each segment depending on its classification (formant transitions are lengthened and the dominating vowel sound and silence periods are compressed in time), and scales each segment by the desired rate (variable rate WSOLA 107) while maintaining the spectral characteristics. The resulting modified signal has a speech waveform with enhanced initial consonants, while having approximately the same time-duration as the original input signal.

[32] Selective frequency band amplification may be applied to enhance the acoustic cues. Non-adaptive modification, however, may create distortions or, in the case of unvoiced fricatives especially, may bias perception in a particular direction. For best emphasis of the perceptually important formants, an adaptive spectral enhancement technique based on the speech spectral estimate is applied. The enhancement filter 103 is based on the linear prediction coefficients. The purpose, however, is not to mask quantization noise as in coding synthesis, but instead to accentuate the formant structure.

[33] The tilt compensator 104 applies tilt compensation after the formant enhancement to reduce negative spectral tilt. For intelligibility, it may be desirable not only to flatten the spectral tilt, but also to amplify the higher frequencies. This is especially true for the distinction of unvoiced fricatives. A high frequency boost reduces the upward spread of the masking effect, in which the stronger lower frequencies mask the weaker upper frequencies. For simplicity, a first order filter is applied.

[34] The adaptive spectral enhancement filter is:

$$H(z) = (1 - \alpha z^{-1}) \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \qquad (3)$$

where, $\gamma_1 = 0.8$, $\gamma_2 = 0.9, \alpha = 0.2$, and $1/A(z)$ is a 10^{th} order all-pole filter which models the speech spectrum. These constants are determined through informal intelligibility testing of confusable word pairs. In the exemplary embodiment the constants remain fixed; however, in variations of the exemplary embodiment they are determined adaptively in order to track the spectral tilt of the current speech frame.

[35] Modification of the phoneme durations is an important part of the enhancement technique. Time-scale modification is commonly performed using overlap-add techniques with constant scaling factor. In some applications, the modification is performed for playback purposes; in other words, the speech signal is stored and then either compressed or expanded for listening, as the user requires. In such applications constraints on speech delay are not strict, allowing arbitrary expansion, and the entire duration of the speech is available a priori. In such cases, processing delays are not of paramount importance, and the waveform can be continuously compressed without

requiring pauses in the output. However, the present invention allows the process to operate at the time of speaking, essentially in real-time. It is therefore necessary to constrain delays, both look-ahead and those caused by signal retardation. Any segment expansions must be compensated by compression of the following segment, in order to provide for speaker-to-speaker interaction. In variable-rate time-scale modification the choice of scaling factor is based on the characteristics of the target speech segment.

[36] First, syllables that are to be expanded /compressed are determined in syllable segmentation 105. In the exemplary embodiment, syllables correspond to the consonant-vowel transitions and the steady-state vowel combinations. The corresponding speech region, as illustrated as boundary 201 in Figure 2, is referred to as the time-scale modification syllable (TSMS). Note, the TSMS only contains quasi-periodic speech. Typically, a TSMS has a time duration between 100 msec to 300 msec. The TSMS does not include the initial features of the consonant such as stop bursts, frication noise, or pre-voicing. Thus, the detection measures that are most appropriate will differ from other time-scale modification techniques, which attempt to locate regions of non-stationarity. In other variations of the exemplary embodiment, other types of speech structures can correspond to a syllable. In general, syllable boundaries can be flexible. For example, the entire vowel sound may or may not be included in the TSMS segment.

[37] Automatic detection of the TSMS is important procedure of the algorithm. Any syllables that are wrongfully identified can lead to distortions and unnaturalness in the output. For example with fast speech, two short syllables may be mistaken for a single syllable, resulting in an undesirable output in which the first syllable is excessively expanded, and the second is almost lost due to full compression. Hence, a robust detection strategy is required. Several methods may be applied to detect TSMS boundaries including the rate of change of spectral parameters (line spectral frequencies (LSFs), cepstral coefficients), rate of change of energy, short-time energy, and cross-correlation measures.

[38] If the look-ahead delay is to be minimized, the most efficient method to locate the TSMS is a cross-correlation measure that it can be obtained directly from WSOLA

synthesis of the previous frame. However, considerable performance improvements (fewer boundary errors and/or distortions in the modified speech) are realized when the TSMS duration is known before its modification begins; hence the reduced complexity advantages cannot be capitalized upon. Both the correlation and energy measures can identify long duration high-energy speech sections of the signal that correspond to voiced portions to be modified. The short-time energy, E_n, of the signal $x(t)$ centered at time $t=n$, is calculated as

$$E_n = \sqrt{\frac{1}{N+1}\sum_{m=-N/2}^{N/2} x^2(n+m)} \qquad (4)$$

where the window length N=20 ms. However, time-domain measures have difficulty discriminating two syllables in a continuous voiced section. TSMS detection is more reliably accomplished using a measure that detects abrupt changes in frequency-domain characteristics, such as the known spectral feature transition rate (SFTR). The SFTR is calculated as the gradient, at time n, between the Line Spectral Frequencies (LSFs), y_l, within the interval $[n \pm M]$. This is given by the equation:

$$SFTR = \sum_{l=1}^{P} \left(g_n^l\right)^2 \qquad (5)$$

where, the gradient of the l^{th} LSF, is

$$g_n^l = \frac{\sum_{m=-M}^{M} m\, y_l(n+m)}{\sum_{m=-M}^{M} m^2}, \quad l = 1,\ldots,P \qquad (6)$$

and P, the order of prediction, is 10. LSFs are calculated every 10ms using a window of 30ms. The SFTR can then be mapped to a value in the range [0, 1], by the function:

$$C_n = \frac{2}{1+e^{-\beta(SFTR)}} - 1 \qquad (7)$$

where, the variable β is set to 20.

[39] In the exemplary embodiment, syllable segmentation is thus performed using a combination of two measures: one that detects variability in the frequency domain and one that identifies the durations of high energy regions. In the exemplary embodiment, the energy contour is chosen instead of the correlation measure because of its reduced complexity. While the SFTR requires the computation of LSFs at every frame, it contributes substantial reliability to the detection measure. Computational savings may be realized if the technique is integrated within a speech encoder. In simplified terms, the boundaries of the TSMS are first estimated by thresholding the energy contour by a predefined value. The SFTR acts as a secondary measure, to reinforce the validity of the initial boundary estimates and to separate syllables occurring within the same high energy region when a large spectral change occurs. Figure 3 illustrates the measures used to detect the syllable to time-scale. An input (speech) signal is processed by lowpass filter 302, energy calculator 304 and energy ratio calculator 306 to provide a ratio of highband to lowband energy that is subsequently utilized for fricative detection. The speech signal is also processed by energy calculator 308 to determine an energy contour. The LSFs from formant emphasis 103 is processed by SFTR 310 to determine a rate of change of LSFs. The energy contour and the rate of change of LSFs are utilized to locate the TSMS boundaries, are shown in Figures 4, 5, and 6. Figure 4 depicts locating the time-scale modification syllable for the word "fin" according to a speech waveform. Boundary 401 corresponds to a TSMS of approximately 175 msec in time duration. Figure 5 depicts locating the time-scale modification syllable for the word "fin" according to an energy contour. Figure 6 depicts locating the time-scale modification syllable for the word "fin" according to a spectral feature transition rate (SFTR).

[40] Since unvoiced fricatives are found to be the least intelligible of the consonants in intelligibility tests previously performed, an additional measure is included to detect frication noise. The energy of fricatives is mainly localized in frequencies beyond the available 4 kHz bandwidth, however, the ratio of energy in the upper half-band to that in the lower half-band is found to be an effective identifying cue. If this ratio lies above a predefined threshold, the segment is identified as a fricative. Further enhancement (amplification, expansion) of these segments is then feasible.

[41] Once the TSMSs have been identified, an appropriate time-scaling factor is dynamically determined by the time scale determinator 106 for each 10 ms-segment of the frame. (A segment is a portion of speech that is processed by a variable-rate scale modification process.) The strategy adopted is to emphasize the formant transitions through time expansion. This effect is then strengthened by compressing the following vowel segment. Hence, the first portion of the TSMS containing the formant transitions is expanded by α_{tr}. The second portion containing the steady-state vowel is compressed by α_{ss}. Fricatives are lengthened by α_{fric}. The scaling factors are defined as follows:

$\alpha < 1$ corresponds to lengthening the time duration of the current segment,
$\alpha > 1$ corresponds to compression, and
$\alpha = 1$ corresponds to no time-scale modification at all.

[42] Time scaling is inversely related to the scaling factor. Typically, $\alpha_{tr} = 1/\alpha_{ss}$; however for increased effect, $\alpha_{tr} < 1/\alpha_{ss}$. Significant changes in time duration, e.g. $\alpha > 3$, may introduce distortions, especially in the case of stop bursts. The factors used in the current implementation are: $\alpha_{tr} = 0.5$, $\alpha_{ss} = 1.8$ and $\alpha_{fric}=0.8$. In low energy regions of the speech, residual delays may be reduced by scaling the corresponding speech regions by the factor $\alpha_{sil} = \min(1.5,\ 1+d/(LF_s))$, where d is the current delay in samples, L is the frame duration and F_s is the sampling rate.

[43] In a variation of the exemplary embodiment of the present invention, the first one third of the TSMS is slowed down and the next two thirds are compressed. However, delay constraints often prevent the full TSMS duration from being known in advance. This limitation depends on the amount of look-ahead delay, D_L, of the algorithm and the speaking rate. Since the ratio of expansion to compression durations is 1:2, the maximum TSMS length, foreseeable before the transition from α_{tr} to α_{ss} may be required, is $1.5*D_L$. If the TSMS duration is greater than $1.5*D_L$, the length of the portion to be expanded is set to a value, $N \geq 0.5*D_L$, which depends on the energy and SFTR characteristics. Compression of the next $2N$ ms then follows; however, this may be interrupted if the energy falls below the threshold during this time.

[44] With D=100 ms, the chosen scaling factors typically result in a total delay less than 150 ms, although delay may peak up to 180 ms very briefly during words containing fricatives. A block diagram of the variable-rate time-scale modification procedure is shown in Figure 7. The underlying technique is WSOLA with an additional facility of accommodating a variable scaling factor. Speech signal 701 (which may be spectrally shaped in accordance with function 101 in Figure 1) is stored in buffer 702 for subsequent processing. The speech signal is variably time-scaled by functions 714 and 710. Function 714 utilizes energy information 715, SFTR 716, and high/low energy ratio information 717 to detect a TSMS and to consequently determine a scaling factor for each region of the speech signal. Depending on the value of the scaling factor, the position of the current and target pointers are adjusted with reposition buffer pointer function 704. With function 706, a search using cross-correlation is then performed to find the segment within a given tolerance of the target position that has maximum similarity to the continuation of the last extracted segment. After each best-match search, the delay is calculated with function 712. This is to ensure that the maximum allowable delay is not exceeded, as well as to determine the current residual delay that may be diminished during future low-energy periods. Since the analysis to determine the desired amount of scaling is performed at a constant rate of time, the scaling factor is updated (with function 710) after each overlap-add operation (function 708) with the value associated with the closest corresponding point in the input signal to provide modified signal 718. With very low energy frames, further compression may take place to reduce the variable residual delay to zero.

[45] Figure 8 shows a flow diagram in accordance with the functional diagram of the exemplary embodiment that is shown in Figure 7. In step 801, a frame of the speech signal is stored into a buffer corresponding to buffer 702) for subsequent processing in accordance with the process shown in Figure 8. (In the exemplary embodiment, the speech signal can correspond to an analog signal or can be digitized by sampling the analog signal and converting the samples into a digital representation to facilitate storing in buffer 702.) The frame is typically of fixed duration of the speech signal (e.g. 20 msec). In step 803 the energy and the SFTR contours (corresponding to energy calculator function 308 and SFTR function 310, respectively) is determined

for further processing in step 805. In step 805, syllable segmentation determines if a TSMS occurs, and if so, the time position of the TSMS. In step 807, if a TSMS is detected and if a consonant-vowel transition occurs (step 808), the corresponding duration speech signal (typically a segment) is time scaled with the scaling factor ($\alpha<1$). However, the corresponding duration of the speech signal is time scaled with a scaling factor ($\alpha>1$) during a steady-state vowel. For other portions of the TSMS, the scaling factor is equal to 1 (in other words, the corresponding speech signal is not time-scaled.) If a TSMS is not detected in step 807, then with step 809 the scaling factor is equal to 1 (no time scaling for the duration of the frame).

[46] In step 811, the frame is processed in accordance with the constituent segments of speech. In the exemplary embodiment, a segment has a time duration of 10 msec. However, other variations of the embodiment can utilize different time durations for specifying a segment. In step 813, the segment is matched with another segment utilizing a cross-correlation and waveform similarity criterion (corresponding to function 706). A best-matched segment within a given tolerance of the target position to the continuation of the extracted segment is determined. (In the exemplary embodiment, the process in step 813 essentially retains the short-term frequency characteristics of the processed speech signal with respect to the inputted speech signal.) In step 815, the scaling factor is adjusted for the next segment of the frame in order to reduce distortion to the processed speech signal.

[47] In step 817, the delay incurred by the segment is calculated. If the delay is greater than a time threshold in step 819, then the scaling factor is adjusted in subsequent segments in order to ensure that the maximum allowable delay is not exceeded in step 821. (Thus, the perceived effect of the real-time characteristics of the processed speech signal is ameliorated.)

[48] In step 823, the segment and the best-matched segment are blended together (corresponding to function 708) by overlapping and added the two segments together, thus providing modified speech signal 718 when all the constituent segments of the frame have been processed in step 825. If the frame has not been completely processed, the buffer pointer is repositioned to correspond to the end of the best-matched segment that was previously determined in step 813. The processed speech

signal is outputted to an external device or to a listener in step 827 when the frame has been completely processed. If the frame has not been completely processed, the buffer pointer is repositioned to the end of the best-matched segment (as determined in step 813) in step 829 so that subsequent segments of the frame can be processed.

[49] Figures 9, 10, 11, and 12 show the original speech waveform for the word "fin", along with the selected scaling factor contour, incurred delay, and the modified output waveform, respectively. The lengthening of the both the "f" frication and the initial parts of the vocalic sections enhances the perception of formant transitions, and hence consonant contrasts. Since the scaling factors are chosen to slightly lengthen the duration of the TSMS, some residual delays are present during the final "n" sound. These are eliminated in the silence period.

[50] Expansion of the initial part of the TSMS often shifts the highest energy peaks from the beginning to the middle of the word. This may affect perception, due to a slower onset of energy. To restore some of the initial energy at onset, the first 50 ms of the TSMS is amplified by a factor of 1.4, with the amplification factor gradually rolling-off in a cosine fashion. A purpose of the amplification is to compensate for reduced onset energy caused by slowing a segment and not to considerably modify the CVR, which can often create a bias shift.

[51] When the above modifications are applied to sentence-length material, the resulting modified speech output sounds highly natural. While the output has a variable delay, the overall duration is the same as the original.

[52] There are two types of delay that are incurred in this algorithm. The look-ahead delay, D_L, is required to estimate the length of each TSMS in order to correctly portion the expansion and compression time durations. This is a fixed delay. The residual delay, D_R, is caused by slowing down speech segments. This is a variable delay. The look-ahead delay and the residual delay are inter-related.

[53] In general, the total delay increases up to $(D_L + N^*\alpha_{tr} + D_R)$ ms, as the formant transitions are lengthened. This delay is reduced, primarily during the remainder of the periodic segment and finally during the following low-energy region. It is not

possible to eliminate 100% of the residual delay D_R during voiced speech if there is to be a smooth continuation at the frame boundaries. This means that the residual delay D_R typically levels out at one pitch period or less until the end of the voiced section is reached.

[54] The best choice for the look-ahead delay D_L depends on the nature of the speech. Ideally, it is advantageous to know the TSMS duration in advance to maximize the modification effect, but still have enough time to reduce the delay during the steady-state portion. This results in minimum residual delays, but the look-ahead delay could be substantial. Alternatively, a minimum look-ahead delay option can be applied, in which the duration of the segment to be expanded is fixed. This means that no look-ahead is required, but the output speech signal may sound unnatural and residual delays will build up if the fixed expansion length frequently exceeds one third of the TSMS duration. If the TSMS duration is underestimated, the modification effect may not reach its full potential. A compromise is to have a method that uses some look-ahead delay, for example 100 ms, and some variable delay.

[55] The present invention combines variable-rate time-scale modification with adaptive spectral enhancement to increase the salience of the perceptually important consonant-vowel formant transitions. This improves the listener's ability to process the acoustic cues and discriminate between sounds. One advantage of this technique over previous methods is that formant transition lengthening is complemented with vowel compression to reinforce the enhanced consonant cues while also preserving the overall speech duration. Hence, the technique can be combined with real-time speech applications.

[56] The drive towards lower speech transmission rates due to the escalating use of wireless communications places high demands on maintaining an acceptable level of quality and intelligibility. The 2.4 kbps Mixed Excitation Linear Prediction (MELP) coder was selected as the Federal Standard for narrowband secure voice coding systems in 1996. A further embodiment of the present invention emphasizes the co-articulation between adjacent phonemes by combining adaptive spectral enhancement with variable-rate time-scale modification (VR-TSM) and is utilized with the MELP coder. Lengthening of the perceptually important formant transitions is

complemented with vowel compression both to reinforce the enhanced acoustic cues and to preserve the overall speech duration. The latter attribute allows the enhancement to be applied in real-time coding applications.

[57] While intelligibility enhancement techniques may be integrated into the coding algorithm, for simplicity and portability to other frameworks, the inventive VR-TSM algorithm is applied as a preprocessor to the MELP coder with the second embodiment. Moreover, other variations of the embodiment may utilize other types of speech coders, including code excited linear prediction (CELP) and its variants, vector sum excitation (VSELP), waveform interpolation (WI), multiband excitation (MBE) and its variants, linear prediction coding (LPC), and pulse code modulation (PCM) and its variants. Since the VR-TSM enhancement technique is applied as a preprocessing block, no alterations to the MELP encoder/decoder itself are necessary. This also allows for emphasis and exaggeration of perceptually important features that are susceptible to coding distortions, to counterbalance modeling deficiencies.

[58] The MELP coding technique is designed to operate on naturally produced speech, which contains familiar spectral and temporal properties, such as a –6dB spectral tilt and, with the exception of pitch doubling and tripling, a relatively smooth variation in pitch during high-energy, quasi-periodic regions. The inventive intelligibility enhancement technique necessarily disrupts some of these characteristics and may produce others that are uncommon in natural speech. Hence, coding of this modified signal may cause some unfavorable effects in the output. Potential distortions in the coded output include high energy glitches during voiced regions, loss of periodicity, loss of pulse peakedness, and irregularities at voiced section onsets.

[59] While both naturalness and the cues for the highly confusable unvoiced fricatives are enhanced with an upward tilt, the emphasis of the high frequency content can create distortions in the coded output. This includes a higher level of hiss during unvoiced speech, "scratchiness" during voiced speech and possibly voicing errors due to the creation of irregular high energy spikes which reduces similarity between pitch periods. On the other hand, formant enhancement, without tilt compensation, reduces the peakedness of pitch pulses. Since MELP synthesis already includes spectral enhancement, additional shaping prior to encoding is unnecessary unless it affects

how well the formants are modeled. While a positive spectral tilt assists the MELP spectral analysis in modeling the higher formants, its accuracy is insufficient to gain intelligibility improvement.

[60] A second potential source of distortion is the search for the best-matched segment in WSOLA synthesis. The criterion of waveform similarity in the speech-domain signal provides a less strict definition for pitch, and as shown in Figure 13, may cause pitch irregularities. Such errors to a single pitch cycle are often unperceivable to the listener, but may be magnified and worsened considerably by low bit-rate coders as shown in Figure 14. In the case depicted, the sudden, irregular shape and duration of one input cycle during a steady, periodic section of speech leads to loss of periodicity and high energy glitches in the MELP output. Glitches may also be produced near the onset of voiced segments if the time-scale modification procedure attempts to overlap-add two segments that are extremely different.

[61] To prevent the above distortions from incurring precautionary measures are included within the intelligibility enhancement preprocessor. The adaptations include the removal of spectral shaping, improved pitch detection and increased time-scale modification constraints. These modifications are motivated by the constraints placed on the input waveform by the MELP coder, and may be unnecessary with other speech coding algorithms such as waveform coding schemes. To prevent irregular pitch cycles, the pitch is estimated every 22.5 ms using the MELP pitch detector prior to WSOLA modification. The interpolated pitch track, $p_{MELP}(i)$, then serves as an additional input to the WSOLA algorithm to guide the selection of the best-matched segment. The pitch as determined using WSOLA, $p_{WSOLA}(i)$, expressed as

$$m\, p_{WSOLA}(n+\tau^{-1}(kL)+\Delta_k) = (1-\alpha)F_L + \Delta_{k-1} - \Delta_k, \quad \begin{matrix} m = 1,2,3,... \\ k = 1,2,3,... \end{matrix} \tag{8}$$

where F_L is the overlap-add segment length, is then constrained during periodic sections to satisfy the condition:

$$p_{MELP}(i) - \delta \le p_{WSOLA}(i) \le p_{MELP}(i) + \delta \ . \tag{9}$$

[62] During transitional regions, especially at voice onsets, interpolation of the MELP pitch is unreliable and hence is not used. During unvoiced speech, the "pitch" is not critical. While this necessarily adds further complexity, a smooth pitch contour is important for low rate parametric coders. Alternatively, a more efficient solution is to integrate a reliable pitch detector within the WSOLA best-match search.

[63] In addition, further constraints are placed on the time-scale modification to avoid the creation of irregularities at voice onsets. A limit is placed on the maximum amount any segment may be expanded ($\alpha \geq 0.5$). No overlap-addition of segments is permitted if the correlation between the best-matched segment and template is below a predefined threshold. This reduces the likelihood of smoothing out voice onsets or repeating an energy burst.

[64] Figure 15 illustrates a functional diagram of the intelligibility enhancement 1512. The speech signal is stored in buffer 1502 for subsequent processing. Syllable segmentation 1504 detects and determines the location of a TSMS. Scaling factor determination function 1506 determines the scaling factor from syllable information form function 1504. If stored speech signal 1504 is characterized by being voiced speech, then pitch detection function 1508 determines pitch characteristics of the speech signal. WSOLA 1510 utilizes scaling information from function 1506 and pitch information from function 1508 in order to process the speech signal. The output of WSOLA is provided to MELPe (which is a variant of the MELP algorithm) coder 1514 for processing in accordance with the corresponding algorithm. (Other variations of the exemplary embodiment can support other types of coders, however.)

[65] Figure 16 is a flow diagram corresponding to the functional diagram that is shown in Figure 15. In step 1601, a frame of the speech signal is stored into a buffer. In step 1603, syllable segmentation (corresponding to function 1504) determines if a TSMS occurs, and if so, the time position of the TSMS. In step 1605, if a TSMS is detected and if a consonant-vowel transition occurs (step 1607), the corresponding duration speech signal (typically a segment) is time scaled with the scaling factor ($\alpha<1$). However, the corresponding duration of the speech signal is time scaled with a scaling factor ($\alpha>1$) during a steady-state vowel. For other portions of the TSMS, the scaling factor is equal to 1 (in other words, the corresponding speech signal is not time-

scaled). If a TSMS is not detected in step 1605, then the scaling factor is equal to 1 in step 1609 (no time scaling for the duration of the frame).

[66] In step 1611, the pitch component of the frame is estimated (corresponding to function 1508). In step 1613, the frame is processed in accordance with the constituent segments of speech. In the exemplary embodiment, a segment has a time duration of 10 msec. If the speech signal corresponding to the segment is voiced as determined by step 1615, then step 1617 determines the best-matched segment using a waveform similarity criterion in conjunction with the pitch characteristics that are determined in step 1611. However, if the speech signal corresponding to the segment is unvoiced, then the best matched segment is determined using the waveform criterion in step 1619 without the necessity of utilizing the pitch information.

[67] If the segment and the best-matched segment are sufficiently correlated as determined in step 1621, then the two segments are overlapped and added in step 1625. However, if the two segments are not sufficiently correlated, the segment is not overlapped and added with the best-matched segment in step 1623. Step 1627 determines if the frame has been completely processed. If so, the enhanced speech signal corresponding to the frame is outputted to a speech coder in step 1629 in order to be appropriately processed in accordance with the associated algorithm of the speech coder. If the frame is not completely processed, then the buffer pointer is repositioned to the segment position in step 1631.

[68] It is to be understood that the above-described embodiment is merely an illustrative principle of the invention and that many variations may be devised by those skilled in the art without departing from the scope of the invention. It is, therefore, intended that such variations be included with the scope of the claims.